数理解析研究所講究録
28巻 1967年 75-86

マルコフ型決定過程(I)

日大 生産工 坂本武司

§1. 平均最適政策

有限個の状態、決定からなる離散型マルコフ型決定過程について述べる。

S個の状態 $1, 2, \cdots, S$ からなる系(System)を考える。離散時点($n=1,2,\cdots\cdots$)で系の状態を観測し、可能な決定の集合 $K=\{k\}=\{1,2,\cdots,K\}$ から決定を選ぶ。もし、n時点で状態 i を観測し、決定 k を行えば、その結果

(1) 系は推移確率 q_{ij}^{k} をもって新しい状態 j に移る。

(2) 状態 i で決定 k を行えば、利得 r_i^k を得る。

即ち.

状態空間: $S=\{1,2,\cdots\cdots,S\}$

決定空間: $K=\{1,2,\cdots\cdots,K\}$

推移確率: q_{ij}^{k} , $\sum_{j=1}^{S} q_{ij}^{k}=1$

利得: r_i^k

f を S から K への関数とし、f の全体を F で表わす。F の要素の系列 $\pi=(f_1, f_2, \cdots, f_n, \cdots)$ を政策と呼ぶ。政策 π を用いるとは初期状態が i_1 であれば、第一期に決定 $f_1(i_1)$、第二期の状態が i_2 であれば $f_2(i_2)$ 以下同様に続けて、第 n 期に $f_n(i_n)$ を行う。$\pi=(f, f, \cdots, f, \cdots)\equiv f^\infty$ を定常政策という。

$f\in F$ に対する利得及び推移確率を次のようにベクトル表示する。

$$r(f)=\left(r_1^{f(1)}, r_2^{f(2)}, \cdots, r_S^{f(S)}\right)' \quad (S\times 1 \text{ ベクトル}) \qquad Q(f)=\left(q_{ij}^{f(i)}\right) \quad (S\times S \text{ 行列})$$

$V_n(i,\pi)$ = 状態 i から出発して、政策 π を用いるときの n 期迄の総期待利得

とする。そのベクトル表示を

$$V_n(\pi)=\left(V_n(1,\pi), \cdots, V_n(S,\pi)\right)' \quad (S\times 1 \text{ ベクトル})$$

とする。そのとき

$$(1\cdot 1) \qquad V_n(\pi)=\sum_{k=0}^{n-1} Q_k(\pi)\, r(f_{n+1})$$

但し $Q_k(\pi)=Q(f_1)\cdots Q(f_k)$, $Q_0(\pi)=I$ (単位行列)

任意の $S\times 1$ ベクトル W に対して作用素 $L(f)$ を

$$(1\cdot 2) \qquad L(f)W = r(f)+Q(f)W$$

で定義する。$L(f)$ は monotone である。即ち、任意のベクトル U, W について、すべての i について $U_i\geqq W_i$ ならば $U\geqq W$、$U\geqq W$ 且つ $U\neq W$ ならば $U>W$ と定義すると、$U\geqq W$ な

らば、$L(f)u \geqq L(f)w$ である。今後上記のようにベクトルの順序を定めておく。$L(f)$ を用いると、

$$V_n(\pi) = L(f_1)L(f_2)\cdots\cdots L(f_{n-1})r(f_n)$$

$V_n(\pi)$ は $n\to\infty$ のとき発散し、1期当りの平均利得 $\frac{1}{n}V_n(\pi)$ は一般に収束しないが、$u(\pi) = \liminf_{n\to\infty} \frac{1}{n} V_n(\pi)$ を考えると $u(f) \geqq u(\pi)$ for all π を満たす定常政策 f^∞ が存在する。但し $u(f) = u(f^\infty) = Q^*(f)r(f)$。また割引率 β を入れると、$\beta \to 1$ のとき、無限過程の総期待利得を最大にする定常政策が存在する。更に、各 n について、$U(n) = \max_{\pi} V_n(\pi)$ とすると、$\lim_{n\to\infty} \frac{1}{n} U(n) = u(f)$ となる定常政策 f^∞ が存在する。そして、この3つの定常政策の平均利得は一致する。そこで、定常政策の中で、平均利得を最大にする政策を決定する algorithm を2一変換及び割引率 β を用いずに求めてみる。

Theorem 1. (i), (ii). 任意の $f \in F$ について、$Q^n(f) \to Q^*(f)$ $(n\to\infty)$ とする。

(1.2) $\quad V_n(f^\infty) = nu(f) + v(f) + \varepsilon(n,f)$

(b). $Q(f)$ のエルゴディク集合の周期の最小公倍数を N とし、$Q_0 \equiv Q^N$ とすると、

(1.3) $\quad V_{nN+m}(f^\infty) = (nN+m)u(f) + v(f) + w(m,f) + \varepsilon(n,f)$

但し、$u(f)$ は

(1.4) $\quad (I - Q(f))\,l = 0 \quad , \quad Q^*(f)u = Q^*(f)r(f)$

の一意の解であり、$v(f)$は

(1.5) $(I - Q(f))v = r(f) - u(f)$ 、 $Q^*(f)v = 0$

の一意の解である。また、$w(nN+m, f) = w(m, f)$ $(n=1,2\cdots)$

$\varepsilon(n, f) \to 0$ $(n \to \infty)$

(2) $G_1(f) \equiv \{ g \mid g \in F,\ Q(g)u(f) > u(f) \}$

$G_2(f) \equiv \{ g \mid g \in F,\ Q(g)u(f) = u(f),\ r(g) + Q(g)v(f) > u(f) + v(f) \}$

$G(f) \equiv G_1(f) \cup G_2(f)$ とする。

$g \in G(f)$ ならば、$u(g) \geqq u(f)$

(3) $G(f) = \Phi$ ならば、任意の $g \in F$ について $u(f) \geqq u(g)$

(証明) (1) (a). $V_n(f^\infty) = \sum_{k=0}^{n-1} Q^k(f) r(f)$

$= (n-1) Q^*(f) r(f) + \sum_{k=0}^{n-1} (Q(f) - Q^*(f))^k r(f)$

$n \to \infty$ のとき、$Q^n(f) \to Q^*(f)$ だから、n が十分大なるとき、

$V_n(f^\infty) = n Q^*(f) r(f) + H(f) r(f) + \varepsilon(n, f)$

$= n u(f) + v(f) + \varepsilon(n, f)$

但し、$v(f) \equiv H(f) r(f)$, $H(f) \equiv (I - Q(f) + Q^*(f))^{-1} - Q^*(f)$

$\varepsilon(n, f) \to 0$ $(n \to \infty)$

(b) $Q^*(f) = \frac{1}{N} Q_0^*(f) \sum_0^{N-1} Q^i(f)$ より

$V_{nN+m}(f^\infty) = (n-1) N Q^*(f) r(f) + [I - (Q_0(f) - Q_0^*(f))]^{-1}$

$\sum_0^{N-1} Q^i(f) r(f) + Q_0^*(f) \sum_0^{m-1} Q^i(f) r(f)$

よって、これより上の結果を得る。

(2). 周期的な場合について示す。仮定より n を十分大きくとると、

$$V_{nN+m+1}(g, f^\infty) - Q(g) W(m, f) - Q(g) \varepsilon(nN+m, f)$$
$$> V_{nN+m+1}(f^\infty) - W(m+1, f) - \varepsilon(nN+m+1, f)$$

即ち、$L(g) V_{nN+m}(f^\infty) > V_{nN+m+1}(f^\infty) + Q(g) W(m, f)$
$$- \varepsilon(nN+m+1, f) + Q(g) \varepsilon(nN+m, f)$$

一般に、任意の自然数 M について、

$$L^M(g) V_{nN+m}(f^\infty) > V_{nN+m+M}(f^\infty) + Q^M(g) W(m, f)$$
$$- W(m+M, f) - \varepsilon(nN+m+M, f) + Q^M(g) \varepsilon(nN+m, f)$$

実際、$M=1$ のとき明らかだから M のとき仮定すると、

$$L^{M+1}(g) V_{nN+m}(f^\infty) \geqq L(g) V_{nN+m+M}(f^\infty) + Q^{M+1}(g) W(m, f)$$
$$- Q(g) W(m+M, f) - Q(g) \varepsilon(nN+m+M, f)$$
$$+ Q^{M+1}(g) \varepsilon(nN+m, f)$$

仮定を用いると、

$$L^{M+1}(g) V_{nN+m}(f^\infty) > V_{nN+m+1}(f^\infty) + Q^{M+1}(g) \cdot W(m, f)$$
$$- W(m+M+1, f) - \varepsilon(nN+m+M+1, f) + Q^{M+1}(g) \varepsilon(nN+m, f)$$

$\min_{\substack{1 \leqq i \leqq S \\ 1 \leqq m \leqq N}} W_i(m, f) = c_1$ 、 $\max_{\substack{1 \leqq i \leqq S \\ 1 \leqq m \leqq N}} W_i(m, f) = c_2$ 、 $\delta = (1, 1, \cdots, 1)'$

とおくと、

$$L^M(g) V_{nN+m}(f^\infty) > V_{nN+m+M}(f^\infty) + (c_1 - c_2) \delta$$
$$- \varepsilon(nN+m+M, f) + Q^M(g) \varepsilon(nN+m, f)$$

$M \to \infty$ とすると、$\varepsilon(nN+m+M, f) \to 0$ 、また $\varepsilon(nN+m, f)$

、Q^Mは有界だから、上式をMで割りM→∞とすると、$u(f) \geq u(f)$を得る。同期的でないときも同様である。

§2. 殆ど最適政策の計算法

割引率β $(0 \leq \beta < 1)$を考えたときの政策πに対する総期待利得を$V_\beta(\pi)$で表わすと、

(2.1) $$V_\beta(\pi) = \sum_{n=0}^{\infty} \beta^n Q_n(\pi) r(f_{n+1}) \quad \text{但し } Q_0(\pi) = I$$

もし、$U(\beta) \equiv V_\beta(\pi^*) \geq V_\beta(\pi)$ for all π が成立つとき、π^*をβ一最適政策と呼ぶ。

また、適当なβ_0が存在して$(0 \leq \beta_0 < 1)$、すべての$\beta_0 \leq \beta < 1$なるβに対して、$V_\beta(\pi^*) \geq V_\beta(\pi)$ for all π が成立つとき、π^*を最適政策と呼ぶ。β－最適ほど最適な定常政策が存在することは知られている。

さて、$U(\beta) - V_\beta(\pi) \to 0$ $(\beta \uparrow 1)$ならばπを殆ど最適(nearly optimal)又は1－最適(1－optimal)と呼ぶ。ここでは、殆ど最適政策を決定するVeinottのalgorithmについて述べる。

まず、準備として D. Blackwell の結果を述べる。

Theorem 1 と同じ手順により

Lemma 1. 任意の$f \in F$について、

(2.2) $$V_\beta(f^\infty) = \frac{u(f)}{1-\beta} + w(f) + \varepsilon(\beta, f), \quad 0 \leq \beta < 1$$

$\varepsilon(\beta, f) \to 0 \quad (\beta \to 1)$

次の集合を定義する。

$F' \equiv \{ f \mid f \in F,\ u(f) \geq u(g) \ \text{all}\ g \in F \}$

$F'' \equiv \{ f \mid f \in F',\ v(f) \geq v(g) \ \text{all}\ g \in F' \}$

(2.2)式の形より

Lemma 2. F''は始ど最適なすべての$f \in F$の集合である。

したがって、F''の要素を決定するalgorithmを作ればよい。定理1より$G(f) = \phi$なら$f \in F'$であるから、F'の中でvを最大にする計算法を求めればよい。

Lemma 3. $f \in F$, $g \in G(f)$ならば: $u(g) > u(f)$ または、$u(g) = u(f)$, $v(g) > v(f)$

(証明) 定理1より、$g \in G(f)$ならば $u(g) \geq u(f)$. また: $u(g) = \lim_{n\to} \frac{1}{n+1} \sum_0^n Q^k(f) r(f) = \lim_{\beta \to 1} (1-\beta) \sum_0^\infty \beta^k Q^k(f) r(f) = \lim_{\beta \to 1} (1-\beta) V_\beta(f^\infty)$ より、十分1に近いβに対して: $V_\beta(g^\infty) > V_\beta(f^\infty)$

$$V_\beta(g^\infty) = \frac{u(g)}{1-\beta} + v(g) + \varepsilon(\beta, g)$$

$$V_\beta(f^\infty) = \frac{u(f)}{1-\beta} + v(f) + \varepsilon(\beta, f)$$

であるから、$u(g) \geq u(f)$ 且つ、$u(g) = u(f)$ ならば $v(g) \geq v(f)$ もし、$(u(g), v(g)) = (u(f), v(f))$ ならば、$g \in G(f)$に矛盾。

次の集合を定義する。

$E(f) = \{ g \mid g \in F,\ Q(g) u(f) = u(f),\ r(g) + Q(g) v(f) = u(f) + v(f) \}$

$E(f)$の定義と(2.3), (1.4)より $g \in E(f)$ ならば $u(g) = u(f)$ となる

したがって $f \in F'$ ならば $E(f) \subset F'$ である。集合 $E(f)$ は計算出来るから $E(f)$ の中で v を最大にすることを考える。

Lemma 4. (Veinott) $f \in F$、$g \in E(f)$ とする。$w(g)$ は

(2.3) $\quad [I - Q(g)]\, w = 0, \quad Q^*(g) w = Q^*(g)(-v(f))$

の一意の解とする。そのとき

(2.4) $\quad v(g) = v(f) + w(g)$

(証明) 一意性は定理2より出る。$g \in E(f)$ より

$$[I - Q(g)]\, v(f) = r(g) - u(f)$$

この式と (2.3) を加えると、

$$[I - Q(g)][v(f) + w(g)] = r(g) - u(f), \quad Q^*(g)[v(f) + w(g)] = 0$$

よって定理2より、(2.4) が出る。

この Lemma より $E(f)$ の中で $w(g)$ を最大にすることは、F を $E(f)$ で $r(f)$ を $-v(f)$ でおきかえれば、F の中で $u(f)$ を最大にするのと同じ形をしていることがわかる。従って $E(f)$ の中で $v(g)$ を最大にするためには、政策反復法を再び使えばよいことになる。

任意の $f \in F$ について、$z(f)$ を (2.5) の一意の解とする。

(2.5) $\quad (I - Q(f))\, z(f) = -v(f), \quad Q^*(f) z(f) = 0$

(2.6) $\quad H(f) \rightleftharpoons \{ g \mid g \in E(f), \; -v(f) + Q(g) z(f) > z(f) \}$

とする。そのとき

Theorem 2. (Veinott)

(1) $f \in F$, $G(f)=\phi$ とする。もし、$v(f) \geq V(g)$ for all $g \in E(f)$ ならば $f \in F''$

(2) $G(f) \cup H(f)=\phi$ ならば $f \in F''$

(3) $g \in H(f)$ ならば $u(g)=u(f)$ であり、更に $v(g)>v(f)$ 又は $v(g)=v(f)$, $z(g)>z(f)$ のいずれかが成立つ。

(証明) (1) $g \in F'-E(f)$ とすると、$G(f)=\phi$, $g \in F'$ より $u(g)=u(f)$。更に、$g \notin E(f)$ より $r(g)+Q(g)v(f)<u(f)+v(f)$

$$V_\beta(g,f^\infty)=\frac{Q(g)u(f)}{1-\beta}+r(g)-Q(g)u(f)+Q(g)v(f)+\varepsilon(\beta,f,g)$$

$$V_\beta(f^\infty)=\frac{u(f)}{1-\beta}+v(f)+\varepsilon(\beta,f)$$

であるから、十分1に近い β に対して、$V_\beta(g,f^\infty)<V_\beta(f^\infty)$ よって $V_\beta(g^\infty)<V_\beta(f^\infty)$ ([1]参照)、即ち $v(f)>v(g)$.

(2) $H(f)=\phi$ であるから、$w(f) \geq w(g)$ for all $g \in E(f)$、然るに $w(f)=0$ であるから、$v(g)=v(f)+w(g) \leqq v(f)+w(f)=v(f)$ for all $g \in E(f)$.

(3) $g \in H(f) \subset E(f)$ より $u(g)=u(f)$、更に、Lemma 3 より、$w(g)>w(f)=0$、又は $w(g)=w(f)=0$, $z(g)>z(f)$.

この定理により、$f_1 \in F$ とし、$f_2, f_3, \cdots$ を $f_{i+1} \in G(f_i) \cup H(f_i)$ であるように選ぶと、$\{u_i(f), v_i(f), z_i(f)\}$ は辞書的順序で増加するから、同じ f_i が2度現われることはない。F は有限集合だから、あるところで $G(f_i) \cup H(f_i)=\phi$ となり、始と最適政策が得られる。

例. $S=\{1,2\}$, $K_1=\{1,2\}$, $K_2=\{1\}$ (K_iは状態iで可能な決定の集合)

$F=\{f,g\}$ 但し $f(i)=1$ $(i=1,2)$

$g(1)=2$, $g(2)=1$

$$r(f)=\begin{pmatrix}3\\-3\end{pmatrix} \quad Q(f)=\begin{pmatrix}\frac{1}{2}&\frac{1}{2}\\\frac{1}{2}&\frac{1}{2}\end{pmatrix} \quad Q^*(f)=\begin{pmatrix}\frac{1}{2}&\frac{1}{2}\\\frac{1}{2}&\frac{1}{2}\end{pmatrix}$$

$$r(g)=\begin{pmatrix}b\\-3\end{pmatrix} \quad Q(g)=\begin{pmatrix}0&1\\\frac{1}{2}&\frac{1}{2}\end{pmatrix} \quad Q^*(g)=\begin{pmatrix}\frac{1}{3}&\frac{2}{3}\\\frac{1}{3}&\frac{2}{3}\end{pmatrix}$$

$$u(f)=\begin{pmatrix}0\\0\end{pmatrix} \quad v(f)=\begin{pmatrix}3\\-3\end{pmatrix} \quad z(f)=\begin{pmatrix}-3\\3\end{pmatrix}$$

$$u(g)=\begin{pmatrix}0\\0\end{pmatrix} \quad v(g)=\begin{pmatrix}4\\-2\end{pmatrix} \quad z(g)=\begin{pmatrix}-\frac{8}{3}\\\frac{4}{3}\end{pmatrix}$$

$G(f)=\phi$, $G(g)=\phi$, $H(f)=g$, $H(g)=\phi$

gが各点ど最適政策.

参考文献

(1) D. Blackwell, Discrete Dynamic Programming, Ann. Math. Statist. 33 (1962) 719−726

(2) B. W. Brown, On the Iterative Method of Dynamic Programming on a Finite Space Discrete Time Markov Processes, Ann. Math. Statist. 36 (1965) 1279−1285

(3) C. Derman, On Sequential Decisions and Markov

chains, Management Science 9 (1962) 16-24

(4) R. A. Howard, Dynamic Programming and Markov Processes, 1960 Technology Press and Wiley

(5) M. Ogawara, A Note on Discrete Markovian Decision Processes, Bull. Math. Statist 11 (1963) 35-42

(6) A. F. Veinott, Jr, On Finding Optimal Policies in Discrete Dynamic Programming, Ann. Math. Statist 37 (1966) 1284-1294